What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions

Hertzum, Morten; Borlund, Pia; Kristoffersen, Kristina Bonde

Published in:
International Journal of Human-Computer Interaction

DOI:
10.1080/10447318.2015.1065691

Publication date:
2015

Document version
Peer reviewed version

Published in *International Journal of Human-Computer Interaction*, vol. 31, no. 9 (2015), pp. 557-570.

What Do Thinking-Aloud Participants Say? A Comparison of Moderated and Unmoderated Usability Sessions

Morten Hertzum

University of Copenhagen, Denmark, email: hertzum@acm.org

Pia Borlund

University of Copenhagen, Denmark, email: sjc900@iva.ku.dk

Kristina B. Kristoffersen

Usertribe, Denmark, email: kristinabonde@gmail.com

Abstract. The value of thinking aloud in usability tests depends on the content of the users' verbalizations. We investigated moderated and unmoderated users' verbalizations during relaxed thinking aloud (i.e., verbalization at levels 1 through 3). Verbalizations of user experience were frequent and mostly relevant to the identification of usability issues. Explanations and redesign proposals were also mostly relevant, but infrequent. The relevance of verbalizations of user experience, explanations, and redesign proposals showed the value of relaxed thinking aloud but did not clarify the tradeoff between rich verbalizations and test reactivity. Action descriptions and system observations – two verbalization categories consistent with both relaxed and classic thinking aloud – were frequent but mainly of low relevance. Across all verbalizations, the positive or negative verbalizations were more often relevant than those without valence. Finally, moderated and unmoderated users made largely similar verbalizations, the main difference being a higher percentage of high-relevance verbalizations by unmoderated users.

Keywords: Thinking aloud, Verbalization, Usability evaluation method, User test, Usability, User experience

1 INTRODUCTION

Evaluation is central to the development of systems that are experienced as usable and satisfying by users. Consequently, reliable and cost-effective usability evaluation methods are in high regard, and considerable research has gone into devising and assessing such methods. Of these methods, the usability test is considered one of the most valuable tools by practitioners (McDonald, Edwards, & Zhao, 2012; Venturi, Troost, & Jokela, 2006) as well as researchers (Dumas & Fox, 2008; Lewis, 2014). In a usability test, an evaluator observes test users' behavior and listens in on their thoughts while they think aloud (Dumas & Loring, 2008; Rubin & Chisnell, 2008). While the observations provide the evaluator with information about what users do, their verbalization of their thoughts adds

information about why they do it and how they experience it. Thus, the users' verbalization is an important source of input for the evaluator in assessing the user experience and determining the usability problems that may hamper system use.

While usability tests have conventionally been conducted in the laboratory with a moderator running the test sessions, unmoderated usability tests have lately become increasingly popular (Liu, Bias, Lease, & Kuipers, 2012). In unmoderated usability tests, the users run the test session themselves, often in their homes, while their behavior and verbalizations are video-recorded for subsequent analysis. The absence of a moderator in unmoderated usability tests – and thereby the absence of a communication partner and of prompting – may affect the user's verbalizations and possibly make them less informative about the user experience. Previous research has investigated the content of verbalizations in moderated usability tests (Bowers & Snyder, 1990; Cooke, 2010; Zhao, McDonald, & Edwards, 2014). In the present study, we compare the content of verbalizations during moderated and unmoderated test sessions. By analyzing what users say while thinking aloud in these two settings, we aim to understand verbalization in usability tests better and to inform the choice of using moderated or unmoderated test sessions.

Different variants of thinking aloud involve the verbalization of different types of information. This study is about relaxed thinking aloud, which constitutes a relaxation of the thinking-aloud protocol recommended by Ericsson and Simon (1993). In relaxed thinking aloud (also known as interactive thinking aloud) the test users are requested to verbalize their thoughts by providing a running commentary of their actions and they are, in moderated tests, prompted for their current thoughts as well as for reflections on their actions. In contrast, classic thinking aloud complies with the recommendations of Ericsson and Simon (1993) and is restricted to the verbalization of information that is already attended by the user to solve the test tasks. Ericsson and Simon's recommendations are often relaxed in practical usability testing (Boren & Ramey, 2000) and we, therefore, focus on relaxed thinking aloud. We note that in moderated test sessions the moderator prompts more during relaxed than classic thinking aloud. The extra prompting aims to elicit verbalizations that are rich in detail about the user's thoughts and reflections. We also note that unmoderated test sessions add reality to the standard recommendation in classic thinking aloud of instructing the user to "act as if you are alone in the room speaking to yourself" (Ericsson & Simon, 1993, p. 376). To the extent that being alone in the room promotes classic thinking aloud, we would expect to see elements of classic thinking aloud in unmoderated sessions. In the following, we review related work, account for our method, present our results, and discuss their implications for research and practice.

2 RELATED WORK

Thinking aloud provides a way for an evaluator to listen in on users' thoughts during a usability test. Previous work has identified different levels of verbalization, investigated their value to usability evaluators, analyzed the content of verbalizations made in moderated tests, and begun to explore unmoderated tests.

2.1 Levels and Value of Verbalization

The primary theoretical framework for understanding thinking aloud is Ericsson and Simon's (1993) seminal work on verbal reporting. They distinguish between verbalizations at three levels. Level 1 is the verbalization of information that is in a user's present focus of attention in verbal form. No intermediate processes are needed to report this information. Level 2 is the verbalization of information that is in a user's present focus of attention but in nonverbal form. To report this information it must be recoded into verbal form. The recoding involves additional processing but does not bring new information into the user's focus of attention. Level 3 is the verbalization of

information that must first be brought into a user's focus of attention. Thus, verbalization at this level influences the user's focus of attention in ways beyond those occasioned by task performance.

Relaxed thinking aloud comprises verbalizations at levels 1 to 3 and, thereby, includes verbalizations in which current thoughts and presently attended information are linked to earlier thoughts and previously attended information. That is, it includes explanations, reasons, and reflections, which may be useful to understanding the user experience. Relaxed thinking aloud reflects how thinking aloud is mostly applied in practical usability testing (Boren & Ramey, 2000) but is known to distort the thought process and change user behavior (Fox, Ericsson, & Best, 2011; Hertzum, Hansen, & Andersen, 2009; Wilson & Schooler, 1991). In contrast, classic thinking aloud consists of instructing users to restrict their verbalizations to levels 1 and 2. While some studies find that classic thinking aloud influences users' perception of time (Hertzum & Holmegaard, 2015) and impairs their performance of spatial tasks (Gilhooly, Fioratou, & Henretty, 2010), most studies find that this variant of thinking aloud provides an accurate record of the thought process without altering it (Ericsson & Simon, 1993; Fox et al., 2011). However, classic thinking aloud excludes the verbalization of explanations, reasons, and reflections and may, therefore, be less informative about the user experience.

Zhao et al. (2014) have compared the value of verbalization to evaluators' detection of usability problems in classic and relaxed thinking aloud. For relaxed thinking aloud, 40 (45%) of a total of 88 usability problems were derived from users' verbalizations, 4 from observing the users, and the remaining 44 from both verbalization and observation. For classic thinking aloud, 12 (22%) of a total of 55 usability problems were derived from verbalizations, 7 from observations, and the remaining 36 from both verbalization and observation. These results suggest that relaxed thinking aloud produces verbalizations more valuable to usability testing. Haak, Jong, and Schellens (2003, 2004) provided further evidence that classic thinking aloud may add little value to usability testing. In one study they found that an average of only 0.5 of a total of 13.9 usability problems identified per test session were derived from verbalizations only (Haak et al., 2003); in another study, it was 1.7 of a total of 9.7 usability problems (Haak et al., 2004).

2.2 Content of Verbalizations

Previous work on the content of users' verbalizations during usability testing has addressed moderated tests and often compared either classic and relaxed thinking aloud or concurrent and retrospective thinking aloud, see Table 1. Consistent with Ericsson and Simon's (1993) framework, the most frequent verbalization categories during classic thinking aloud tend to be description of action and reading of on-screen text, whereas the most frequent categories for relaxed thinking aloud also include explanation of actions and expression of feelings about the system. The less frequent verbalization categories show that during classic thinking aloud users do not exclusively verbalize at levels 1 and 2 but occasionally also at level 3. Cooke (2010) combined classic thinking aloud with eye tracking and found that the users' verbalizations were verified by their eye movements 80% of the time. This verification of the verbalizations was not broken down on verbalization categories and, therefore, does not tell whether some categories were verified to a larger extent than others. The five other studies in Table 1 have not attempted to assess the accuracy of the verbalizations by, for example, comparing them with independent data. Zhao and McDonald (2010) found a higher number of verbalizations during relaxed than classic thinking aloud for seven of their ten verbalization categories. Conversely, Zhao et al. (2014) found no differences in the number of verbalizations between classic and relaxed thinking aloud.

INSERT TABLE 1 ABOUT HERE

The most common approach to usability testing is that the users think aloud while solving tasks with the system (concurrent thinking aloud). Alternatively, they may first work silently with the system to solve the tasks and afterward think aloud while viewing a video-recording of their use of the system (retrospective thinking aloud). In comparing concurrent and retrospective thinking aloud, Bowers and Snyder (1990, p. 1273) found that "Subjects in the concurrent protocol condition seem to be attending to the experimental tasks and give little thought to the comments they are giving. The retrospective subjects, on the other hand, can give their full attention to the verbalization and in doing so give richer information." This finding suggests that the richer verbalizations that motivate the use of relaxed thinking aloud may also be obtained with retrospective thinking aloud. In addition, retrospective thinking aloud have been found to produce more verbalizations that were relevant to the detection of usability problems than concurrent thinking aloud (McDonald, Zhao, & Edwards, 2013; Ohnemus & Biers, 1993).

The study by Haak, Jong, and Schellens (2006) differs from the others in that the users worked in teams of two or three people who communicated with each other about the tasks and the system. It has been argued that communication with a teammate (i.e., constructive interaction) is more natural than thinking aloud, yet produces similar information (Nielsen, 1993). However, the verbalization categories in the study by Haak et al. (2006) differ markedly from those in the five other studies. Haak et al. (2006) did not report the relevance of the verbalizations in the different categories to the detection of usability problems.

2.3 Unmoderated Tests

Whereas the studies mentioned above concern usability tests in which the users and a human moderator were co-located (though possibly separated by a one-way mirror), usability tests may also be conducted without the presence of a moderator (Liu et al., 2012; Nelson & Stavrou, 2011). In unmoderated tests, the users receive written instructions and conduct the test sessions on their own, often in their homes and typically recruited via crowdsourcing. The crowdsourcing approach to usability testing has become increasingly popular because it offers easy access to users with diverse backgrounds, speedy completion of test sessions, and low cost (Liu et al., 2012). At the same time, Liu et al. (2012, p. 7) warn that "the quantity of feedback from a single crowdsourcing participant is much lower than the quantity of feedback from a single lab test participant." Because unmoderated usability testing is a rather recent possibility, the amount of research is however limited.

In one approach to unmoderated usability testing the users receive instruction to think aloud and their verbalizations and interactions with the tested system are video-recorded for subsequent analysis by an evaluator (e.g., Hertzum, Molich, & Jacobsen, 2014). It is this approach we focus on in the present paper, but we note that unmoderated usability testing also includes variants in which the identification of usability problems is left to the users (Bruun, Gull, Hofmeister, & Stage, 2009). Hertzum et al. (2014) found no difference in the total number of usability problems identified by evaluators who analyzed moderated and unmoderated sessions from a usability test of the same system. They found, however, that evaluators who analyzed moderated test sessions identified more problems that were rated critical or severe than evaluators who analyzed unmoderated test sessions. These findings suggest that the quality of the moderated and unmoderated users' verbalizations and system interactions was comparable with respect to the information needed to identify usability problems but that the unmoderated sessions produced less convincing information about the severity of the problems.

3 METHOD

To investigate verbalization in usability tests, we conducted a between-subjects experiment comparing participants' verbalizations in moderated and unmoderated test sessions.

3.1 Test Participants

We recruited 14 participants for the experiment. The participants were screened for their ability to think aloud by a usability professional, who remained unaware of which participants were assigned to which condition. All participants passed this screening. The participants were an average of 28.7 (moderated) and 29.3 (unmoderated) years of age. For both moderated and unmoderated sessions the gender distribution was 4 female and 3 male participants. This age and gender distribution matched that of the primary user group of the tested website. All participants were Danes and reported using the Internet daily.

3.2 Website and Test Tasks

The tested website, gaffa.dk, was a Danish news site for music. In addition to news, reviews, and feature articles, the site, for example, provided playlists, videos, and photos of contemporary music events and an archive of Danish music history since 1983. At the time of testing (November-December, 2012), the Gaffa website had about 145,000 unique visitors a month.

The test participants received five test tasks, which represented common uses of the website, see Table 2. While Task 1 was open-ended and partly aimed at ensuring an unintimidating session onset, the four other tasks were goal-directed. Tasks 2 to 5 consisted of one instance of each of the four task types identified by Spool, Scanlon, Schroeder, Snyder, and DeAngelo (1999): fact tasks, comparison-of-fact tasks, judgment tasks, and comparison-of-judgment tasks.

INSERT TABLE 2 ABOUT HERE

3.3 Procedure

The participants took part in either a moderated or an unmoderated test session. Before they were invited for the test session, participants completed a trial session during which they tried to think aloud. The trial session was conducted online and used to screen the participants with respect to their ability to think aloud.

Each *moderated test session* involved a single participant. These sessions were conducted by a usability professional from Snitker, who received guidelines about how to moderate the sessions but remained unaware of the specific focus of our study. The moderator welcomed the participants to the laboratory, explained what was going to happen, and instructed them in thinking aloud. This instruction consisted of the following statement: “While you solve the tasks, you are to think aloud. That is, you are to explain what you are in doubt about on the website, what you like, what you dislike, and so forth.” The moderator remained in the room with the participants while they solved the tasks and probed them for information when the participants fell silent for some time, when they became visibly surprised without verbalizing why, and when they had completed a task. Also, when the participants asked the moderator a question, such as “Can I do this?”, the moderator answered with another question, such as “What do you think will happen if you do?” The sessions were video recorded for later analysis.

The *unmoderated test sessions* were conducted remotely by means of the crowdsourcing service brugertest.nu. For these sessions the test tasks, a link to the Gaffa website, and a user profile were uploaded to brugertest.nu. On the basis of the user profile, which specified the required age and gender distribution of the participants, matching users from the brugertest.nu database were invited to the test and, if they accepted, performed the test session online. Participants could perform the test at any time by logging on to the brugertest.nu website, which provided them with the link to the

Gaffa website, the test tasks, and a reminder to turn on the video-recording of the session and to think aloud during the session. When the participants felt they had completed a task, they proceeded to the next task. After completing the entire test session, the participants uploaded the video-recording of the session to brugertest.nu. In these sessions there was no human moderator to probe for information about the participant's thoughts or otherwise facilitate the progress of the session. We presume the participants performed the sessions at home but the remote nature of the unmoderated sessions means that we do not know the specific settings in which these sessions were performed.

At the end of *both moderated and unmoderated tests sessions*, the participants were asked to rate their experience of the Gaffa website with respect to its usefulness and how easy, pleasant, and frustrating it had been to use the site. The exact wording of the four items was:

- Gaffa.dk was useful for solving the tasks
- It was easy to use Gaffa.dk
- Gaffa.dk was pleasant to use
- It was a frustrating experience to use Gaffa.dk

All four items were rated on five-point rating scales with the end points 'strongly disagree' (1) and 'strongly agree' (5). The four items were read aloud to participants by the moderator (moderated sessions) or presented in writing as an additional task after the five tasks that involved using the Gaffa website (unmoderated sessions). We acknowledge that the difference in presentation format may have subtly affected participants' responses and would, in hindsight, have preferred to present the four items in writing in all sessions.

The video-recordings of the test sessions showed the screen with gaffa.dk and thereby visualized the participants' interactions with the website, such as scrolling, clicks, and other mouse movements. Of specific relevance to this study, the recordings also gave the participants' verbalizations. The test sessions lasted an average of 26 minutes. Of these 26 minutes, an average of 20.7 (moderated) and 21.9 (unmoderated) minutes were spent on the five test tasks. Our analysis concerns the time spent on the five tasks.

3.4 Data Analysis

We analyzed the content of the verbalizations by categorizing them according to three classifications: topic, valence, and relevance (see Table 3). The topic classification was devised on the basis of previous studies of the content of verbalizations during usability testing (Bowers & Snyder, 1990; Cooke, 2010; Haak et al., 2006; McDonald et al., 2013; Zhao & McDonald, 2010; Zhao et al., 2014), supplemented with reading the verbalizations from two test sessions. Table 3 shows which categories have been used in which previous studies. The second classification concerned the valence of the verbalizations and simply distinguished between positive and negative verbalizations. We included this classification to be able to analyze how the test sessions balanced a focus on usability problems against one on positive usability issues. The third classification was adopted from Zhao and McDonald (2010) and consisted of assessing the relevance of the verbalizations to usability evaluation. We distinguished between low, medium, and high relevance.

INSERT TABLE 3 ABOUT HERE

In preparation for the categorization, the test sessions were transcribed verbatim. For the moderated test sessions, all utterances by the moderator were marked up to distinguish them from the

participants' verbalizations. The participants' verbalizations were segmented into sentential units. We note that in spoken language sentential units may contain restarts, be left unfinished, and otherwise be messy in structure and content. In the following we refer to a sentential unit as a verbalization. The 14 test sessions comprised 1928 verbalizations by the participants.

For each of the three classifications, the categorization of the verbalizations involved four steps. First, a training set of 203 verbalizations was categorized by two of the authors independently. The training set consisted of a randomly selected moderated test session and a randomly selected unmoderated test session. Each verbalization was assigned either to one of the classification categories or to an 'other' category. Second, all disagreements in the authors' categorizations of the training set were discussed to reach consensus about the categorization of the verbalizations and to create a shared understanding of the classifications. Third, the remaining 1725 verbalizations (12 test sessions) were categorized by the two authors independently. Fourth, all disagreements in the authors' categorizations of these verbalizations were discussed and a consensus was reached.

The topic and valence categorizations were made by the first and second author, the relevance categorization by the first and third author, who was a practicing usability professional with years of experience in analyzing data from usability tests. Prior to categorizing the relevance of the verbalizations, the third author had analyzed the test sessions to identify the usability problems and positive usability issues encountered by the participants. This analysis strengthened the basis for her relevance categorization.

We assessed the reliability of our categorizations of the verbalizations using Cohen's (1960) kappa. For the 1725 non-training verbalizations the kappa values of the agreement between the two authors were .60, .69, and .61 for the topic, valence, and relevance classifications, respectively. These values met the criterion that kappa values of a minimum of .60 indicate satisfactory reliability (Lazar, Feng, & Hochheiser, 2010).

4 RESULTS

The statistical analyses comparing the test conditions (moderated, unmoderated) were made using t-tests with statistical significance set at the level of .05. As a safeguard against any violations of the prerequisites of parametric t-tests we reran the analyses using non-parametric Mann-Whitney U tests and found the exact same significant effects. We note that the statistical analyses comparing the test conditions had modest power. We, thus, cannot rule out that insufficient sample size masked some differences between moderated and unmoderated sessions.

4.1 Control Variables

The participants rated the Gaffa website as useful, easy to use, pleasant, and non-frustrating, see Table 4. For all four of these variables the median response was the same for moderated and unmoderated participants. This non-difference in user experience precluded that the comparison of the verbalizations in the moderated and unmoderated sessions was confounded by an unintended difference in the participants' general experience of the Gaffa website.

INSERT TABLE 4 ABOUT HERE

4.2 Number of Verbalizations and Words

The participants made an average of 157 (moderated) and 118 (unmoderated) verbalizations, see Table 5. There was no effect of test condition on the number of verbalizations, $t(12) = 1.91$, $p = .080$.

To investigate the possibility that differences in the duration of sessions might have masked an effect of test condition, we also analyzed verbalizations per minute. There was a significant effect of test condition on verbalizations per minute, $t(12) = 2.90$, $p = .013$, with moderated participants making more verbalizations per minute than unmoderated participants.

We counted the number of words in the verbalizations to analyze whether words and verbalizations were affected similarly by test condition. There was no effect of test condition on number of words, $t(12) = -1.58$, $p = .14$, and words per minute, $t(12) = -1.94$, $p = .077$. There was, however, a significant effect of test condition on words per verbalization, $t(12) = -4.46$, $p = .001$, with unmoderated participants making longer verbalizations than moderated participants. The larger number of verbalizations per minute by moderated participants combined with the shorter length of these verbalizations suggests a more conversational kind of thinking aloud during the sessions with a moderator. In these sessions, the moderator made an average of 61 verbalizations ($SD = 12$) and spoke an average of 26 words per minute ($SD = 6.29$). Examples of verbalizations by the moderator included "Okay. Was that… Do you think that was easy or difficult to find?" and reading aloud the next test task.

INSERT TABLE 5 ABOUT HERE

4.3 Topic

Table 6 shows the percentage of verbalizations in each category of the topic classification. The four largest categories were action description, system observation, user experience, and 'other'.

Action description included verbalizations such as "Okay, I click on articles" and "[I] try the same again. Try searching up in, eh". These verbalizations accounted for what participants were doing, without providing much information about why they did it or how they experienced doing it. On multiple occasions participants appeared to resort to action description when they did not know what else to say. Action description constituted 19% (moderated) and 27% (unmoderated) of the verbalizations.

System observation was similarly frequent. It constituted 18% (moderated) and 24% (unmoderated) of the verbalizations. Examples of system observation included "And, then it froze again, here" and "It is an ad, I think. That I am watching. It is also down here. You cannot click on it. It is an ad, yes". This category of verbalizations described the website and only provided little information about what participants were doing or how they experienced the website. Like action descriptions, system observation was more about the observable world outside participants' heads than about participants' cognitive processes.

User experience constituted 19% of the verbalizations. Examples of these verbalizations included "I, I find it a little difficult to understand that this is one of the main articles. It is wildly uninteresting" and "He-he, that was easy". Verbalizations about user experience revealed aspects of what went on inside participants' heads, that is, how they perceived using the website. Words such as "annoying", "confusing", "cool", "difficult", "easy", "exciting", "fine", "interesting", "messy", and their opposites were common in the user-experience verbalizations.

The 'other' category consisted of the verbalizations that did not relate to any of the six topic categories. They often had little content. An example of these verbalizations was "Yes, you know… eh".

INSERT TABLE 6 ABOUT HERE

The three remaining categories in the topic classification were much less frequent, each accounting for only 3-5% (moderated) and 2-4% (unmoderated) of the verbalizations. These three categories were explanation (e.g., "And then I worked out that I should not look up here because that is only the last five. I have to go down here. And then click"), redesign proposal (e.g., "And it could also be cool if you had, if I had a search function so I could take one of my, some of my favorite bands and type them in and simply get a, get a list of when they play"), and domain knowledge (e.g., "Well, I mostly use Facebook for everything to do with music"). The verbalizations in these categories were reflections, in which participants talked about why they used the website the way they did, how they thought it could be improved, and how they otherwise obtained music information.

The topic classification captured 68% (moderated) and 77% (unmoderated) of the verbalizations; the remaining verbalizations were categorized as 'other'. We found a significant effect of test condition on the percentage of 'other' verbalizations, $t(12) = 2.91$, $p = .013$, indicating that a larger part of the verbalizations were outside the topic categories for moderated than unmoderated participants. A number of these additional 'other' verbalizations were brief acknowledgements, such as "Uh" and "Yes", of the moderator's verbalizations. For the six non-'other' categories, there was no effect of test condition on the percentage of verbalizations in a category (action description: $t(12) = -2.07$, $p = .061$, explanation: $t(12) = 1.56$, $p = .15$, system observation: $t(12) = -1.38$, $p = .19$, redesign proposal: $t(12) = -0.53$, $p = .60$, domain knowledge: $t(12) = 1.68$, $p = .12$, user experience: $t(12) = -0.06$, $p = .95$).

4.4 Valence

With respect to valence, 19% (moderated) and 24% (unmoderated) of the verbalizations were positive and another 20% (moderated) and 18% (unmoderated) were negative, see Table 7. That is, positive and negative verbalizations were about equally frequent. The negative verbalizations included "I hate all these ads" and "Okay. Hmm, then I would expect that the menu from just before was here. But it is not". The positive verbalizations also spanned a range from very explicit to more tempered, for example "Fuck this is easy. And I can just search on artist. So, we write: Tina Dickow" and "Well, I actually think it is okay with those ads. If there has to be ads then these are fine. As long as they do not make noise". As many as 61% (moderated) and 58% (unmoderated) of the verbalizations were neither positive nor negative. Many of these 'other' verbalizations were purely descriptive, such as "And here I get it as a PDF, eh". We found no effect of test condition on the percentages of positive, negative, and other verbalizations, ts(12) = -1.04, 0.47, 0.59, respectively (all ps > .32).

INSERT TABLE 7 ABOUT HERE

Figure 1 shows how the positive and negative verbalizations were distributed on the topic categories. Four results stand out. First, nearly all user-experience verbalizations were either positive or negative. For moderated participants the distribution was an average of 14.71 (positive), 12.57 (negative), and 2.00 (other). For unmoderated participants it was 11.57 (positive), 7.57 (negative), and 2.14 (other). Only 7% (moderated) and 10% (unmoderated) of the verbalizations in the user-experience category were neither positive nor negative. The second-smallest percentage of neither positive nor negative verbalizations was for redesign proposals (moderated) with 24% and explanations (unmoderated) with 33%.

Second, action descriptions were mostly neither positive nor negative. As much as 76% (moderated) and 78% (unmoderated) of the verbalizations in the action-description category were neither positive nor negative. System observations were somewhat similar with 51% (moderated) and 50% (unmoderated) of verbalizations being neither positive nor negative.

Third, redesign proposals were rarely positively worded. A moderated participant made an average of 0.29 positive, 3.71 negative, and 1.29 other redesign-proposal verbalizations. For an unmoderated participant the averages were 0.14 (positive), 2.14 (negative), and 2.14 (other). Thus, most redesign proposals took their starting point in participants' negative perception of the current design of the website. Explanations displayed a similar picture with an average of 8% (moderated) and 17% (unmoderated) positive verbalizations.

Fourth, the distribution of positive and negative verbalizations across the topic categories was similar for moderated and unmoderated participants.

INSERT FIGURE 1 ABOUT HERE

4.5 Relevance

About half of the verbalizations were of low relevance to the identification of usability issues and about one quarter were of medium relevance, see Table 8. These verbalizations included "This one is very calm. There is, well, there is basically just what I expect, eh" (low) and "Ehm, you could perhaps be a bit worried to, eh, to miss something if you do not choose the right one of those three" (medium). We found no effect of test condition on the percentage of low-relevance, $t(12) = -0.52$, $p = .62$, and medium-relevance, $t(12) = 1.21$, $p = .25$, verbalizations. There was, however, a significant effect of test condition on the percentage of high-relevance verbalizations, $t(12) = -4.12$, $p = .001$, with more high-relevance verbalizations by unmoderated than moderated participants. Examples of high-relevance verbalizations were "Well, what I really want to do now is just to go to Google and search because this is, eh" and "I just try to click on it and see what happens. It needs some time for downloading… Well, this takes quite some time". There was also a significant effect of test condition on 'other' verbalizations, $t(12) = 6.64$, $p = .001$, with more verbalizations not captured by the relevance classification in the moderated than unmoderated sessions.

INSERT TABLE 8 ABOUT HERE

Figure 2 shows how relevance was distributed across the topic categories. Four results stand out. First, redesign proposals had the highest percentage of verbalizations of medium and high relevance. An average of 89% (moderated) and 87% (unmoderated) of the redesign-proposal verbalizations were of medium or high relevance. Redesign proposals were followed by user-experience verbalizations with an average of 67% (moderated) and 60% (unmoderated) verbalizations of medium and high relevance and by explanations with 56% (moderated) and 67% (unmoderated). That is, two of the three infrequent topic categories yielded high percentages of verbalizations relevant to the identification of usability issues.

Second, action description and system observation had high levels of verbalizations of low relevance. For action description an average of 61% (moderated) and 65% (unmoderated) of verbalizations were of low relevance, for system observation it was 61% (moderated) and 54% (unmoderated). That is, two of the frequent topic categories yielded mainly low-relevance verbalizations and only the

third of the frequent topic categories, user experience, yielded mainly verbalizations relevant to the identification of usability issues.

Third, the additional high-relevance verbalizations for unmoderated, compared to moderated, participants were distributed across multiple topic categories, thereby suggesting that the difference owed to the unmoderated test condition in general. Explanations were the only category with a higher percentage of high-relevance verbalizations for moderated (21%) than unmoderated (17%) participants.

Fourth, 98% of the verbalizations in the 'other' category of the relevance classification were also in the 'other' category of the topic classification, and predominantly for moderated participants. This result reinforces that the moderation introduced a number of verbalizations that acknowledged the moderator's verbalizations but had no direct relation to the identification of usability issues.

INSERT FIGURE 2 ABOUT HERE

Figure 3 shows the relationship between the relevance and valence classifications. Verbalizations of high and medium relevance were mostly either positive or negative. As much as 71% (moderated) and 67% (unmoderated) of the high-relevance verbalizations were valenced. Of the medium-relevance verbalizations it was 70% (moderated) and 58% (unmoderated). Conversely, most low-relevance verbalizations were neither positive nor negative. We tested this association between relevance and valence using the Goodman and Kruskal tau test, which indicates how much errors in the prediction of one variable are reduced given information about another variable (Costner, 1965). Knowing whether or not verbalizations were valenced reduced errors in the prediction of their relevance by 11% (p = .001, N = 1099) for moderated participants and 10% (p = .001, N = 829) for unmoderated participants.

INSERT FIGURE 3 ABOUT HERE

4.6 Number of Usability Issues

To qualify the interpretation of the verbalizations, especially the relevance classification, Table 9 shows the average number of usability problems and positive usability issues encountered by a participant. We acknowledge that the identification of usability problems and positive usability issues was made by a single person and may, therefore, be subject to an evaluator effect (Hertzum et al., 2014). There was no difference between moderated and unmoderated participants in the number of usability issues (i.e., the sum of usability problems and positive usability issues), $t(12) = 1.64$, $p = .13$.

INSERT TABLE 9 ABOUT HERE

5 DISCUSSION

The participants in the moderated sessions spoke an average of 110 words per minute. In addition, the moderator spoke 26 words per minute. This corresponds to near continuous speech for most of the session. Contrary to the common advice that moderators must remind test users to think aloud

(e.g., Dumas & Loring, 2008), it appears that a moderator was not required to uphold this rate of verbalization: Without a moderator to prompt them for their thoughts, the unmoderated participants spoke at an average rate of 132 words per minute. Previous studies provide no information about the number of words spoken. However, Zhao et al. (2014) provide the time on task and the number of verbalizations. Their participants made an average of 21 verbalizations per minute during relaxed thinking aloud (and 19 during classic thinking aloud). This is substantially more than in our study (cf. Table 5) but the difference may, in part, be an artifact of differences in how participants' thinking aloud was segmented into verbalizations. Zhao et al. (2014) report that their participants needed a reminder to 'keep talking' about once a minute. Our participants' ability to think aloud near continuously may be a result of self-selection among the people who agreed to take part in the study or it may be a result of the instructions and think-aloud screening administered prior to the test session. With the substantial amount of thinking aloud by moderated as well as unmoderated participants, we contend that, at least for relaxed thinking aloud, the main issue is not to keep users talking but to ensure relevant content of their verbalizations.

5.1 What Do Thinking-Aloud Participants Say?

A total of 38% (moderated) and 44% (unmoderated) of the verbalizations were of medium or high relevance to the identification of usability issues. These percentages support previous studies finding that users' verbalizations during relaxed thinking aloud add value to usability tests (Zhao et al., 2014). The relevant verbalizations were, however, unevenly distributed across the topic categories.

Action description was a frequent verbalization topic in our study as well as in previous studies (Table 1). Notably, this category of verbalizations is frequent in both classic and relaxed thinking aloud (Zhao & McDonald, 2010; Zhao et al., 2014). Our finding that action descriptions were predominantly neither positive nor negative appears consistent with a descriptive verbalization category that fits into both classic and relaxed thinking aloud. The limitation of action descriptions is that the majority of these verbalizations were of low relevance to the identification of usability issues. This finding accords with McDonald et al. (2013), who found that none of the action descriptions in their study were relevant to problem detection but noted that action descriptions may still provide useful insights into users' task solving process. System observation was also a frequent verbalization topic in our study, more so than in previous studies. For example, Cooke (2010) reported a percentage of system observations about half as big as we found for our participants. System observations were positively or negatively worded about as often as they displayed no valence, thereby indicating that although many of these verbalizations were merely descriptive, about equally many contained an element of assessment. In spite of this element, system observations resembled action descriptions in their high percentage of low-relevance verbalizations.

Apart from action description and system observation, the verbalization categories go beyond classic thinking aloud. The four additional verbalization categories are, thus, the rationale for preferring relaxed over classic thinking aloud. Previous studies showed that participants instructed to do classic thinking aloud also made verbalizations in the additional categories (e.g., Zhao & McDonald, 2010). Thus, while these categories may formally be specific to relaxed thinking aloud, they occur, in practice, during both types of thinking aloud. The most frequent of the additional categories was user experience with 19% of the verbalizations, about evenly divided between positive and negative. Zhao et al. (2014) found a somewhat higher percentage of user-experience verbalizations for relaxed thinking aloud but in their study negative user-experience verbalizations were five times as frequent as positive. A possible explanation for the dominance of negative over positive user-experience verbalizations is the quality of the tested website: Zhao et al. (2014) found roughly three times as many usability problems in their website as we did in ours. An important quality of the user-experience verbalizations was that the majority of them were of medium or high relevance. This finding speaks in favor of including user-experience verbalizations in usability tests. The relevance of

user-experience verbalizations to usability-problem identification is supported by Zhao and McDonald (2010), who found that 24% of these verbalizations were relevant, compared to only 12% of the total set of verbalizations.

Explanations, redesign proposals, and domain knowledge accounted for a total of 12% (moderated) and 8% (unmoderated) of the verbalizations. That is, the total percentage of verbalizations in these three categories was lower than the percentage of verbalizations in each of the categories action description, system observation, and user experience. In previous studies of relaxed thinking aloud, the percentage of explanations alone was 11% and 16% in Zhao and McDonald (2010) and Zhao et al. (2014), respectively. Notably, Cooke (2010) reported a percentage of explanations during classic thinking aloud comparable to our findings during relaxed thinking aloud. These previous studies suggest that small changes in instructions and probing may suffice to increase the frequency of, at least, explanations. Such an increase would be desirable because the majority of explanations and, especially, redesign proposals were relevant to the identification of usability issues. Consistent with our findings, redesign proposals were consistently found relevant but infrequent in previous studies (Bowers & Snyder, 1990; McDonald et al., 2013; Zhao & McDonald, 2010). In a study of how website developers assess the outcome of usability tests, Hornbæk and Frøkjær (2005) provided further support for the value of redesign proposals. The developers considered redesign proposals more useful to their work than mere descriptions of usability problems. It should, however, be noted that the redesign proposals were made by usability evaluators rather than verbalized by users.

The valence of verbalizations was associated, though only moderately, with their relevance in that positive or negative verbalizations were more often relevant to the identification of usability issues than verbalizations that were neither positive nor negative. This association contributes to explaining how relaxed thinking aloud adds value to usability tests compared to classic thinking aloud, which at least in principle does not contain verbalizations of users' positive or negative reaction to the tested website. While valenced verbalizations may facilitate usability analysis by making usability professionals aware of users' feelings, assessments, and other positive or negative reactions, it is less clear whether they are similarly aware of how relaxed thinking aloud may subtly change users' behavior (McDonald et al., 2012; Nørgaard & Hornbæk, 2006). We note that verbalizations of no direct relevance to the identification of usability issues may augment usability evaluators' understanding of users' behavior and, thereby, enable that additional usability implications can be drawn from the behavior. This way, low-relevance verbalizations may indirectly facilitate usability analysis in a manner beyond our relevance classification of the verbalizations.

Greenberg and Buxton (2008) have argued that usability testing is, at times, harmful because its focus on the identification of usability problems may quash innovative design ideas by enumerating their imperfections rather than stimulating their maturation. On this basis we find it encouraging that participants made roughly similar numbers of positive and negative verbalizations. While this may in part reflect the quality of the tested website, it shows that positive website qualities can be verbalized – if instructions and prompting aim to elicit both positive and negative verbalizations. It is subsequently up to the usability evaluators to accurately report positive and negative verbalizations in their analysis and reporting of the test results.

5.2 Comparing Moderated and Unmoderated Test Sessions

The moderated and unmoderated sessions were similar in many respects. The distribution of verbalizations on topic categories was similar, except for a higher percentage of 'other' verbalizations by moderated participants. This difference suggests that the moderator's presence produced a number of verbalizations that were mainly directed toward the process of communication with the moderator and largely unrelated to the specific activity of testing the website. The distribution of verbalizations on valence categories was also similar for moderated and unmoderated participants, just as the number of positive and negative verbalizations in most topic categories was similar for

moderated and unmoderated participants. Finally, the number of more and less relevant verbalizations was in most topic categories roughly similar for moderated and unmoderated participants. The main difference between the two test conditions was that the unmoderated participants made a higher percentage of high-relevance verbalizations, an average of 21% compared to 11% for the moderated participants. The difference in high-relevance verbalizations did, however, not lead to the identification of more usability issues (Table 9). We conjecture that the unmoderated participants' high-relevance verbalizations contained more duplication and thus constituted a stronger set of evidence for the same usability issues. These findings speak against the warning by Liu et al. (2012) that an unmoderated participant provides much less usability information than a moderated participant. In addition, unmoderated tests are low cost and provide easy and speedy access to users with diverse backgrounds (Liu et al., 2012; Nelson & Stavrou, 2011).

5.3 Implications for Research and Practice

Our study has implications for research as well as practice. First, the content of users' verbalizations has not been studied in detail. More research is needed to investigate, for example, the relationship between users' verbalizations and actions and the impact of their verbalizations on the outcome of usability tests. It is also worth noting that our topic classification, devised on the basis of previous studies, failed to capture 32% (moderated) and 23% (unmoderated) of the verbalizations. What topics were these verbalizations about? One of several candidate categories is verbalizations about the test tasks: learning their content, revisiting tasks to obtain specific pieces of information, and announcing the answers to tasks.

Second, the highest percentages of relevant verbalizations were for user experience and the other topic categories that include level 3 verbalizations. This finding reiterates the need for understanding the tradeoffs involved in eliciting level 3 verbalizations from users. While studies in cognitive psychology provide valuable insights about this issue (e.g., Fox et al., 2011), there is a scarcity of studies that address it from a usability-testing perspective, with Boren and Ramey (2000) and Zhao et al. (2014) as important exceptions. The amount of verbalization by the moderator warrants specific analyses of these verbalizations. Such analyses could establish an empirical basis for recommending prompts that elicit useful user verbalizations, including improved advice about how to avoid level 3 verbalizations when they are unwanted (previous studies have shown that they occur with some frequency during classic thinking aloud).

Third, unmoderated usability testing is a rather recent possibility and the studies conducted so far have yielded mixed results. Studies agree that unmoderated testing is easy, quick, low cost, and thereby of practical interest, but disagree about the information obtained through unmoderated testing. Although we found similar verbalization content for moderated and unmoderated participants and Hertzum et al. (2014) reported no difference in the number of usability problems identified, others have reported that unmoderated tests yield less usability information (Liu et al., 2012) and are restricted to run "a very basic online usability study" (Nelson & Stavrou, 2011, p. 1083). More research is needed.

Fourth, usability testing is sensitive to cultural differences. For example, Clemmensen, Hertzum, Hornbæk, Shi, and Yammiyavar (2009) reported that, depending on their cultural background, users may respond differently to the requirement to verbalize their thoughts and moderators may construe users' verbalizations differently. Yet, the existing studies of verbalization in usability tests have been restricted to Europe and North America. This bias may limit the applicability of the findings to these regions. Cross-cultural studies of the content of verbalizations and their impact on the outcome of usability tests in different regions of the world are welcome.

In terms of practical implications, we recommend that usability professionals:

(1) focus less on keeping users talking and more on obtaining relevant verbalizations, notably this presupposes clarity about which verbalization topics are relevant to the test

(2) take valence as a hint about relevance and maintain the balance between positive and negative verbalizations in their analysis

(3) prompt more for explanations and redesign proposals, when wanted, because they tend to be relevant and may otherwise be infrequent

(4) run some sessions, especially in cross-cultural usability tests, with retrospective or no thinking aloud to experience the tradeoffs between rich verbalizations and test reactivity

(5) consider the use of unmoderated usability tests, which on the basis of this study appear to yield verbalizations similar in content and more often of high relevance compared to moderated tests.

5.4 Limitations

Five limitations should be remembered in interpreting the results of this study. First, we studied the content of verbalizations during relaxed thinking aloud, which takes multiple forms. Formally, relaxed thinking aloud is defined by the presence of verbalizations at level 3 but the content of these verbalizations depends on the instructions and, for moderated sessions, the prompting. Because there is no widely accepted set of instructions and prompting guidelines, relaxed thinking aloud is, in practice, vaguely defined. We acknowledge that others may practice it differently from how it was done in this study. Second, we reiterate that the modest sample size of our study with seven participants in each test condition may mask differences between moderated and unmoderated participants. Thus, the absence of significant differences in most of our statistical tests is not strong evidence of a real absence of differences. Our strongest argument for claiming that moderated and unmoderated sessions are in many respects alike is the many similarities in our qualitative analysis of the verbalizations. Third, in the categorization of the verbalizations the authors were not blind to whether the verbalizations were made by a moderated or unmoderated participant. Thus, our tacit expectations about how the test condition might influence participants' verbalizations may have influenced the categorization. Fourth, we have not linked the content analysis of the verbalizations directly to a process of identifying the usability issues on the website. The absence of a direct link weakens the implications of our analysis for the practice of usability testing. However, the relevance classification provides a surrogate link, which was strengthened by the third author's previous analysis of the sessions to identify the usability issues encountered by the participants. Fifth, we studied verbalizations from the evaluation of a single website. It would be valuable to replicate the study with different websites. These websites should, preferably, span a range from low to high usability, as determined by an assessment independent of the usability test that delivers the verbalizations for the study.

6 CONCLUSION

The value of having users verbalize their thoughts during a usability test depends on the content of their verbalizations. We have investigated verbalizations during relaxed thinking aloud and found that action description, system observation, and user experience were the most frequent topic categories. Whereas action descriptions and system observations tended to display no valence and be of low relevance, the user-experience verbalizations were predominantly either positive or negative and relevant to the identification of usability issues. Of the less frequent topic categories, most of the explanations and redesign proposals were relevant, thereby suggesting that usability professionals may want to prompt for more of these verbalizations. Topic categories such as user experience, explanations, and redesign proposals go beyond classic thinking aloud. Their relevance shows the value of relaxed thinking aloud but without clarifying the tradeoff between richer

verbalizations and test reactivity. Across all verbalizations those displaying a valence were more often relevant to the identification of usability issues than the verbalizations that were neither positive nor negative. This finding adds to the value of relaxed thinking aloud, which allows for verbalizations of users' feelings, assessments, and other positive or negative reactions to the tested system. Our study included moderated as well as unmoderated test sessions. The verbalizations made by moderated and unmoderated participants were similar in content, the main difference being a higher percentage of high-relevance verbalizations by unmoderated participants. On this basis, we recommend that usability professionals consider the use of unmoderated usability tests, at least as a supplement to conventional moderated tests.

ACKNOWLEDGEMENTS

We are grateful to Vanessa Goedhart Henriksen from brugertest.nu for screening the test participants for their ability to think aloud and to Birna Dahl from Snitker for conducting the moderated test sessions. We thank Annika Olsen for transcribing the participants' verbalizations. In the interest of full disclosure, we note that at the time of the study, the third author was an intern in Snitker. Special thanks are due to the test participants.

REFERENCES

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication, 43*(3), 261-278.

Bowers, V. A., & Snyder, H. L. (1990). Concurrent versus retrospective verbal protocols for comparing window usability. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 1270-1274). Santa Monica, CA: HFS Press.

Bruun, A., Gull, P., Hofmeister, L., & Stage, J. (2009). Let your users do the testing: A comparison of three remote asynchronous usability testing methods. In *Proceedings of the CHI 2009 Conference on Human Factors in Computing Systems* (pp. 1619-1628). New York: ACM Press.

Clemmensen, T., Hertzum, M., Hornbæk, K., Shi, Q., & Yammiyavar, P. (2009). Cultural cognition in usability evaluation. *Interacting with Computers, 21*(3), 212-220.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cooke, L. (2010). Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions on Professional Communication, 53*(3), 202-215.

Costner, H. L. (1965). Criteria for measures of association. *American Sociological Review, 30*(3), 341-353.

Dumas, J. S., & Fox, J. E. (2008). Usability testing: Current practice and future directions. In A. Sears & J. A. Jacko (Eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications* (Second ed., pp. 1129-1149). New York: Erlbaum.

Dumas, J. S., & Loring, B. (2008). *Moderating usability tests: Principles & practices for interacting*. Burlington, MA: Morgan Kaufmann.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data. Revised edition*. Cambridge, MA: MIT Press.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137*(2), 316-344.

Gilhooly, K. J., Fioratou, E., & Henretty, N. (2010). Verbalization and problem solving: Insight and spatial factors. *British Journal of Psychology, 101*(1), 81-93.

Greenberg, S., & Buxton, B. (2008). Usability evaluation considered harmful (some of the time). In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 111-120). New York: ACM Press.

Haak, M. J. v. d., Jong, M. D. T. d., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology, 22*(5), 339-351.

Haak, M. J. v. d., Jong, M. D. T. d., & Schellens, P. J. (2004). Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. *Interacting with Computers, 16*(6), 1153-1170.

Haak, M. J. v. d., Jong, M. D. T. d., & Schellens, P. J. (2006). Constructive interaction: An analysis of verbal interaction in a usability setting. *IEEE Transactions on Professional Communication, 49*(4), 311-324.

Hertzum, M., Hansen, K. D., & Andersen, H. H. K. (2009). Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology, 28*(2), 165-181.

Hertzum, M., & Holmegaard, K. D. (2015). Thinking aloud influences perceived time. *Human Factors, 57*(1), 101-109.

Hertzum, M., Molich, R., & Jacobsen, N. E. (2014). What you get is what you see: Revisiting the evaluator effect in usability tests. *Behaviour & Information Technology, 33*(2), 143-161.

Hornbæk, K., & Frøkjær, E. (2005). Comparing usability problems and redesign proposals as input to practical systems development. In *Proceedings of the CHI 2005 Conference on Human Factors in Computing Systems* (pp. 391-400). New York: ACM Press.

Lazar, J., Feng, J. H., & Hochheiser, H. (2010). *Research methods in human-computer interaction*. Chichester, UK: Wiley.

Lewis, J. R. (2014). Usability: Lessons learned ... and yet to be learned. *International Journal of Human-Computer Interaction, 30*(9), 663-684.

Liu, D., Bias, R. G., Lease, M., & Kuipers, R. (2012). Crowdsourcing for usability testing. In *ASIST2012: Proceedings of the American Society for Information Science and Technology* (pp. 1-10). Hoboken, NJ: Wiley.

McDonald, S., Edwards, H. M., & Zhao, T. (2012). Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication, 55*(1), 2-19.

McDonald, S., Zhao, T., & Edwards, H. M. (2013). Dual verbal elicitation: The complementary use of concurrent and retrospective reporting within a usability test. *International Journal of Human-Computer Interaction, 29*(10), 647-660.

Nelson, E. T., & Stavrou, A. (2011). Advantages and disadvantages of remote asynchronous usability testing using Amazon mechanical turk. In *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting* (pp. 1080-1084). Santa Monica, CA: HFES Press.

Nielsen, J. (1993). *Usability engineering*. Boston, MA: Academic Press.

Nørgaard, M., & Hornbæk, K. (2006). What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of the Sixth DIS Conference on Designing Interactive Systems* (pp. 209-218). New York: ACM Press.

Ohnemus, K. R., & Biers, D. W. (1993). Retrospective versus concurrent thinking-out-loud in usability testing. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* (pp. 1127-1131). Santa Monica, CA: HFES.

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests* (Second ed.). Indianapolis, IN: Wiley.

Spool, J. M., Scanlon, T., Schroeder, W., Snyder, C., & DeAngelo, T. (1999). *Web site usability: A designer's guide*. San Francisco, CA: Morgan Kaufmann.

Venturi, G., Troost, J., & Jokela, T. (2006). People, organizations, and processes: An inquiry into the adoption of user-centred design in industry. *International Journal of Human-Computer Interaction, 21*(2), 219-238.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology, 60*(2), 181-192.

Zhao, T., & McDonald, S. (2010). Keep talking: An analysis of participant utterances gathered using two concurrent think-aloud methods. In *Proceedings of the NordiCHI2010 Conference on Human-Computer Interaction* (pp. 581-590). New York: ACM Press.

Zhao, T., McDonald, S., & Edwards, H. M. (2014). The impact of two different think-aloud instructions in a usability test: A case of just following orders? *Behaviour & Information Technology, 33*(2), 163-183.

Table 1. Previous studies of the content of verbalizations during thinking aloud

Study	Thinking aloud	Participants	Findings
Bowers & Snyder (1990)	Concurrent (classic) vs retrospective	48 (between groups)	Concurrent participants gave more verbalizations stating what they were doing and reading out loud words on the screen. Retrospective participants gave more explanations of their actions and more comments on the system design
Cooke (2010)	Classic	10	The most frequent verbalization categories were reading from the screen and description of action
Haak et al. (2006)	Constructive interaction	40 (in 17 teams)	The most frequent verbalization categories were suggestions for how to perform tasks, agreement with a teammate's verbalizations, description of action, and evaluation of action
McDonald et al. (2013)	Concurrent (classic) vs retrospective	10 (within groups)	During concurrent thinking aloud the most frequent verbalization categories were description of action and evaluation of the results of action. During retrospective thinking aloud the most frequent verbalization categories were indications of problems with the system and of the user experience caused by the system
Zhao & McDonald (2010)	Classic vs relaxed	20 (within groups)	For both classic and relaxed thinking aloud the most frequent verbalization categories were description of action, reading of on-screen text, and evaluation of action
Zhao et al. (2014)	Classic vs relaxed	16 (between groups)	For both classic and relaxed participants the most frequent verbalization categories were description of action and expression of negative feelings about the system

Table 2. Test tasks

No.	Task description
1	Go to gaffa.dk and spend two minutes getting an impression of the website by browsing it and exploring what you consider interesting.
2	When does [a named artist] give a concert in Aarhus in the Scandinavian Congress Center?
3	Who has received the highest rating by Gaffa for their latest album – “Machine Gun Kelly” or “Mikael Simpson”?
4	Find the online version of this month’s Gaffa Magazine, describe your experience of the magazine and explain what you like and dislike.
5	Find two interesting articles on the Gaffa website and explain which you like the more, and why.

Table 3. The three classifications used in categorizing the verbalizations

Classification	Category definitions
Topic	
Action description [a,b,c,d,e,f]	Verbalizations describing what participants are doing, trying to do or did, including the reading out loud of text and links on the screen
Explanation [a,b,d,e,f]	Verbalizations explaining why participants act the way they do; explanations may be given before, during, or after actions are performed
System observation [a,b,c,d,e]	Verbalizations making observations about the tested system, including description of features and visual layout
Redesign proposal [a,c,d,e]	Verbalizations offering recommendations on how to improve the system or resolve experienced difficulties
Domain knowledge [d]	Verbalizations mentioning knowledge or past experience with similar tasks or systems, including knowledge of the tested system domain
User experience [d,e,f]	Verbalizations expressing positive or negative feelings and experience resulting from the use of the system
Valence	
Positive	Verbalizations conveying approval, satisfaction, and other positive reactions – experiential as well as utilitarian – to the system
Negative	Verbalizations conveying disapproval, dissatisfaction, and other negative reactions – experiential as well as utilitarian – to the system
Relevance	
Low	Verbalizations contributing negligibly to an analysis of the usability problems and positive usability issues experienced by participants
Medium	Verbalizations supporting the identification of a usability problem or positive usability issue but not themselves sufficient evidence to report it
High	Verbalizations decisive, by themselves, to the identification of a usability problem or positive usability issue

Note. The topic categories have previously been used by: [a] Bowers & Snyder (1990), [b] Cooke (2010), [c] Haak et al. (2006), [d] McDonald et al. (2013), [e] Zhao & McDonald (2010), [f] Zhao et al. (2014).

Table 4. Ratings of the user experience, *N* = 14 participants

Rating scale	Moderated		Unmoderated	
	Median	*Range*	*Median*	*Range*
Useful	4	3-5	4	2-5
Easy to use	4	3-5	4	3-5
Pleasant to use	4	3-5	4	3-5
Frustrating to use	2	1-5	2	1-2

Note: all ratings were made on a five-point scale from 1 (strongly disagree) to 5 (strongly agree)

Table 5. Verbalizations and words by participants, *N* = 14 participants

	Moderated		Unmoderated	
	Mean	*SD*	*Mean*	*SD*
Verbalizations	157	47	118	25
Verbalizations per minute *	7.82	2.14	5.38	0.61
Words	2292	834	2869	485
Words per minute	110	23	132	19
Words per verbalization ***	14.75	3.82	24.70	4.51

* $p < .05$, *** $p < .001$

Table 6. Percent of verbalizations in the categories of the topic classification, *N* = 14 participants

	Moderated		Unmoderated	
	Mean	*SD*	*Mean*	*SD*
Action description	19	6.05	27	8.37
Explanation	5	6.02	2	1.54
System observation	18	5.52	24	9.34
Redesign proposal	3	1.50	4	3.61
Domain knowledge	4	2.12	2	1.51
User experience	19	5.11	19	10.71
Other *	32	3.20	23	8.01

Note: Due to rounding the mean percentages for the unmoderated sessions sum to 101. * $p < .05$

Table 7. Percent of verbalizations in the categories of the valence classification, *N* = 14 participants

	Moderated		Unmoderated	
	Mean	*SD*	*Mean*	*SD*
Positive	19	5.71	24	10.26
Negative	20	5.65	18	7.67
Other	61	6.76	58	11.45

Table 8. Percent of verbalizations in the categories of the relevance classification, *N* = 14 participants

	Moderated		Unmoderated	
	Mean	*SD*	*Mean*	*SD*
Low	53	7.56	56	10.63
Medium	27	6.37	23	6.26
High ***	11	3.05	21	5.89
Other ***	9	3.47	1	0.30

Note: Due to rounding the mean percentages for the unmoderated sessions sum to 101. *** $p < .001$

Table 9. Usability issues encountered by participants, *N* = 14 participants

	Moderated		Unmoderated	
	Mean	*SD*	*Mean*	*SD*
Usability problems	8.00	3.00	7.00	3.27
Positive usability issues	6.00	1.63	4.43	3.64

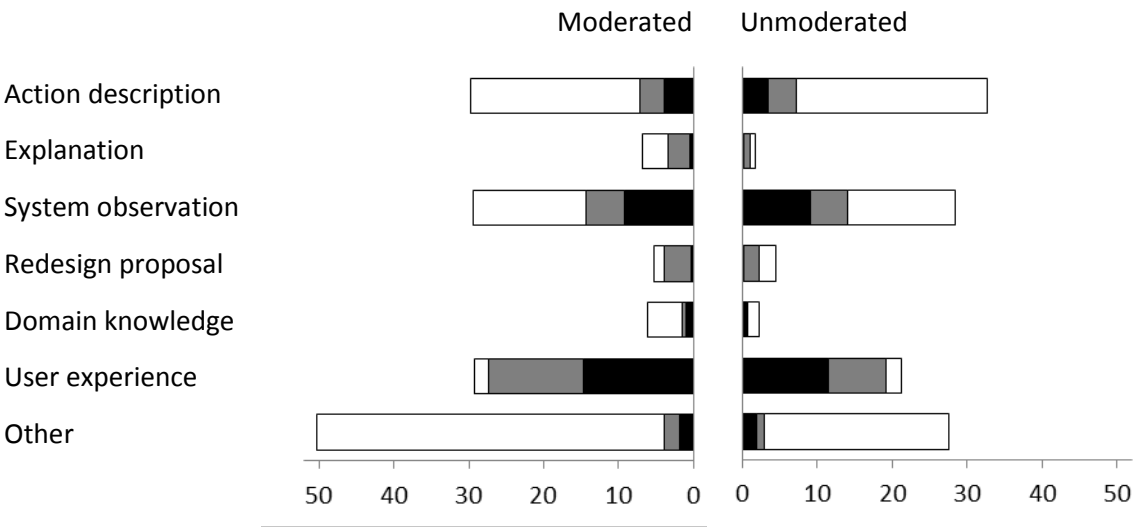

Figure 1. Average number of verbalizations made by participants, divided onto positive (black), negative (grey), and other (white), *N* = 14 participants

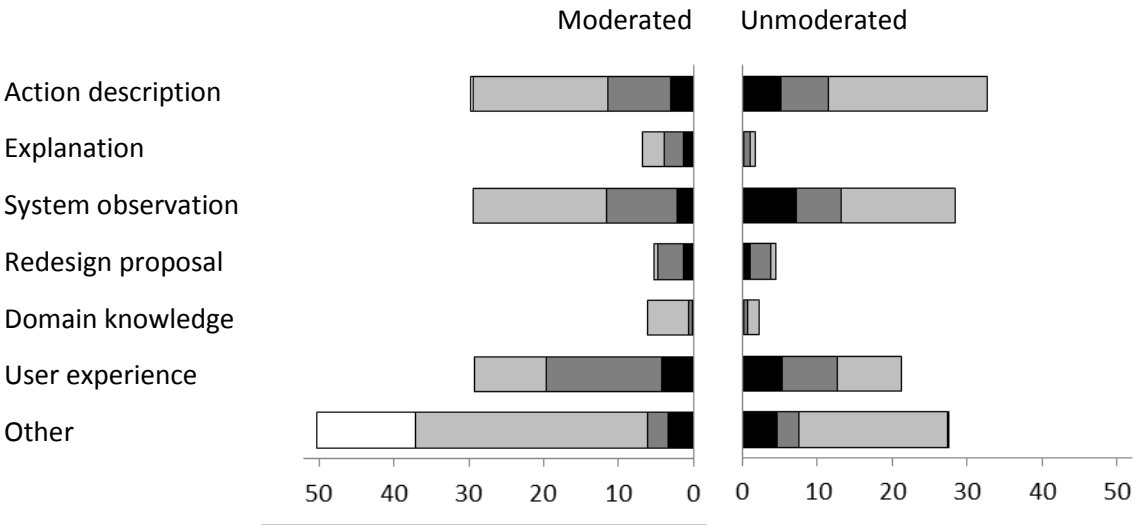

Figure 2. Average number of verbalizations made by participants, divided onto high relevance (black), medium relevance (dark grey), low relevance (light grey), and other (white), *N* = 14 participants

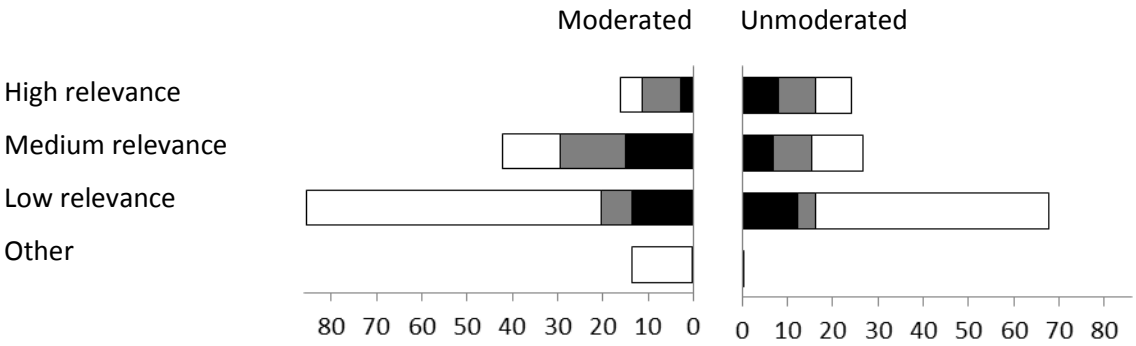

Figure 3. Average number of verbalizations made by participants, divided onto positive (black), negative (grey), and other (white), *N* = 14 participants